BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 9: e183199
doi: 10.3897/biss.9.183199

OPEN ACCESS

Conference Abstract

Untangling Attribution in Biodiversity Data Records

Arturo H. Ariño[‡], Piluca Álvarez Fidalgo[§], Ana Amezcua[‡], Berta Caballero-López[|], Francisco J. Cabrero-Sañudo[¶], Angel Chaves[‡], Marina González Cristóbal[§], Sandra Grzechnik[¶], María Imas[‡], Irene Lobato-Vila[|], Sergio Montagud[#], Mercedes París[§], Adrián Sánchez Albert[§], Manuel Sánchez Ruiz[§], Celia Santos[§], Robert J. Wilson[§], David Galicia[‡]

‡ University of Navarra, Pamplona, Spain
§ Museo Nacional de Ciencias Naturales (MNCN-CSIC), Madrid, Spain
| Museu de Ciències Naturals de Barcelona (MCNB), Barcelona, Spain
¶ Complutense University of Madrid, Madrid, Spain
# University of Valencia, Valencia, Spain

Corresponding author: Arturo H. Ariño (artarip@unav.es)

Received: 22 Dec 2025 | Published: 23 Dec 2025

Citation: Ariño AH, Álvarez Fidalgo P, Amezcua A, Caballero-López B, Cabrero-Sañudo FJ, Chaves A, González Cristóbal M, Grzechnik S, Imas M, Lobato-Vila I, Montagud S, París M, Sánchez Albert A, Sánchez Ruiz M, Santos C, Wilson RJ, Galicia D (2025) Untangling Attribution in Biodiversity Data Records. Biodiversity Information Science and Standards 9: e183199. https://doi.org/10.3897/biss.9.183199

Abstract

The exactness, fitness-for-purpose (FFP) and reliability of primary biodiversity data can be enhanced by additional data beyond the basic taxon-location-date triad (Hill et al. 2010). Often, the only available data are the labels in legacy specimen collections. The digitization process is most efficient if all available information can be collected at once in a single event of specimen handling, rather than in separate phases. There is, however, a compromise between producing a faster catalogue for immediate use and housekeeping, and an accurate, wider-FFP database where all data have been thoroughly checked.

Recognizing potential sources of error at digitization time may help making choices. During a dataset integration procedure, the quality and reliability of the data capture was analyzed. The dataset consisted of transcribed label data of over 58K pinned insects of

© Ariño A et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

agriculturally-relevant groups in XXth-century collections at six institutions in Spain, that resulted in almost 6000 collector strings. But collector names could be

1. unidentified;
2. misread;
3. ambiguous,
4. duplicated under variants; or
5. misplaced or misattributed to/from another entity, e.g. a location.

This resulted in a high entropy level where one collector could be databased in multiple ways, artificially inflating the corresponding catalogues. The entropy was much higher in collections where collectors contributed few specimens, which is the case for university-based collections.

By using simple indexing and cross-referencing techniques, the roster of names was significantly reduced, but full disambiguation of collectors required mining ancillary sources and consulting with people with long-standing knowledge of the collections. Overall, 42% of collector names were in error, resulting in excess entropy.

METHODS

Specimen data were collated from the TETTRIS INC-STEP Project of Spanish pollinators complemented by some agriculturally-relevant groups in six collections deposited at five academic and research institutions in Madrid, Barcelona, Valencia and Pamplona (see Suppl. material 1 for full details). MCNB, MNCN, MUVHN were chiefly historical and created by researchers over long careers, while MZNA-R, MZNA-Z, UCME came mainly from academic coursework activities.

Our procedures agreed with the overall strategy of Groom et al. (2022), while devising a specific workflow. Disambiguation and attribution included a number of steps (Suppl. material 1). First, names were normalized to facilitate grouping and sorting (7 steps). Then, ambiguous names were (whenever possible) attributed to actual persons by internal checks (clustering of names, matching localities and dates), consultation with external references (e.g. student lists), or consultation with collection curators (5 steps).

Names were finally given a four-level identity qualification resulting from the disambiguation exercise (see Suppl. material 1). For further analyses, “very low” and “low” levels were considered a poor attribution, while “medium” and “full” levels were deemed good attribution.

RESULTS

The disambiguation exercise reduced collector names from 5939 to 3473 (a 41.5% decrease). The total entropy, measured as Shannon’s H’, decreased by 8% while Simpson’s dominance D increased by 29%, as expected (merging names created larger collections for some collectors). However, there were differences among collections. The three academic collections had both higher diversity and higher diversity reduction than

the historical collections (Fig. 1). Historical collections tended to be much more concentrated (30 specimens per collector, 25% single-specimen collectors) than the academic collections (8 and 61%, respectively).

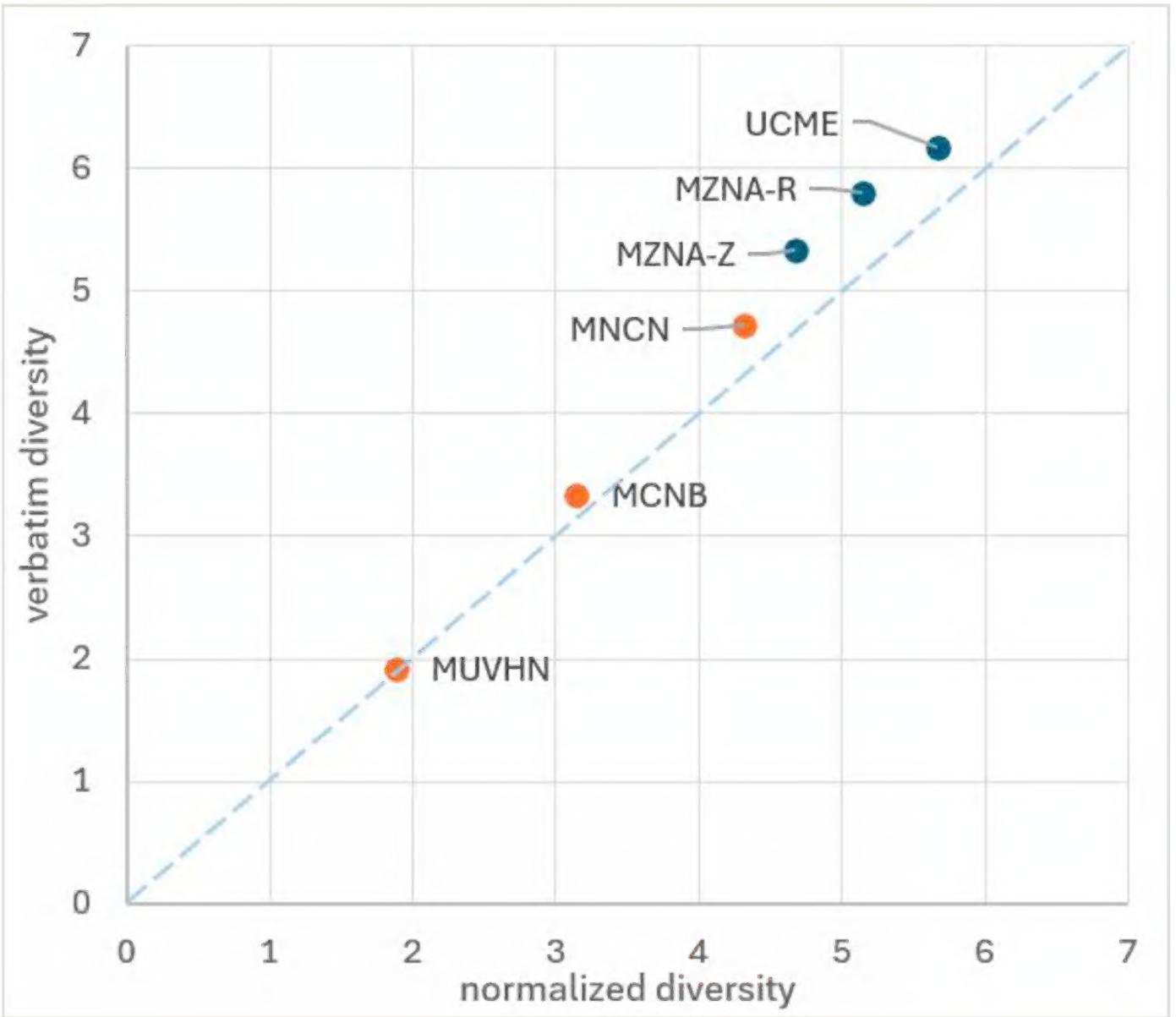

Figure 1.

Entropy of the collections as grouped by the verbatim collector names and by the normalized, attributed names. Blue: mainly academic collections; orange: mainly historical collections. © 2025 A.H. Ariño, CC BY-SA 4.0

These differences can be tracked to how the collections were formed. Historical collections tended to be well documented and created by researchers spanning longer careers, while academic collections were often linked to works created during coursework. Fig. 2 shows the career spans of collectors. The abrupt end of collection at the turn of the century for academic collections can be linked to the introduction of restrictive legislation in Spain (a mandatory reduction of course loads, and unaffordable permission requirements).

Full disambiguation often requires manual, time-consuming verifications against a variety of external sources. Automated disambiguation procedures elevated good identification from an initial 21% to 47% in MZNA-R, but it was not possible to use manual checking against sources for this dataset. In contrast, the similar MZNA-Z collection could be

checked against coursework lists, which resulted in a 80% good attribution rate (Fig. 3). Overall, 57% of the collector names across collections had a good attribution.

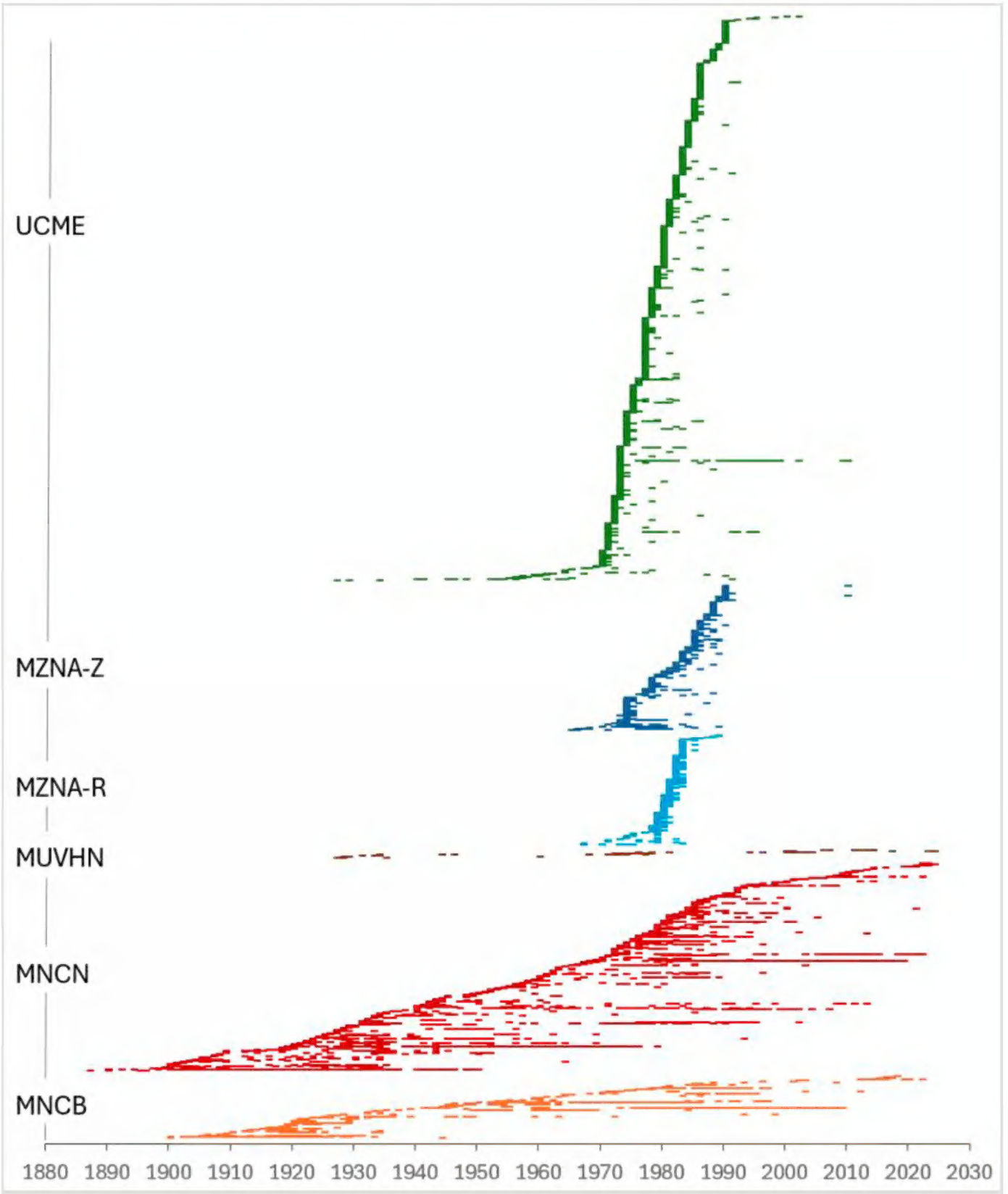

Figure 2.

Collection spans of collectors. Reds: Historical collections; blue/greens: academic collections. Each horizontal line is a collector. © 2025 A.H. Ariño, CC BY-SA 4.0

CONCLUSION

Doing disambiguation exercises by people with contemporary knowledge of the collections appears to be critical to avoid losing much information from the collections. Curators have an invaluable knowledge of the collections, could locate relevant

documents about them, and are the ones able to reduce collection entropy by at least 32%. Their early retirement would negate a significant way to disambiguate names that left little or no trace in the literature record.

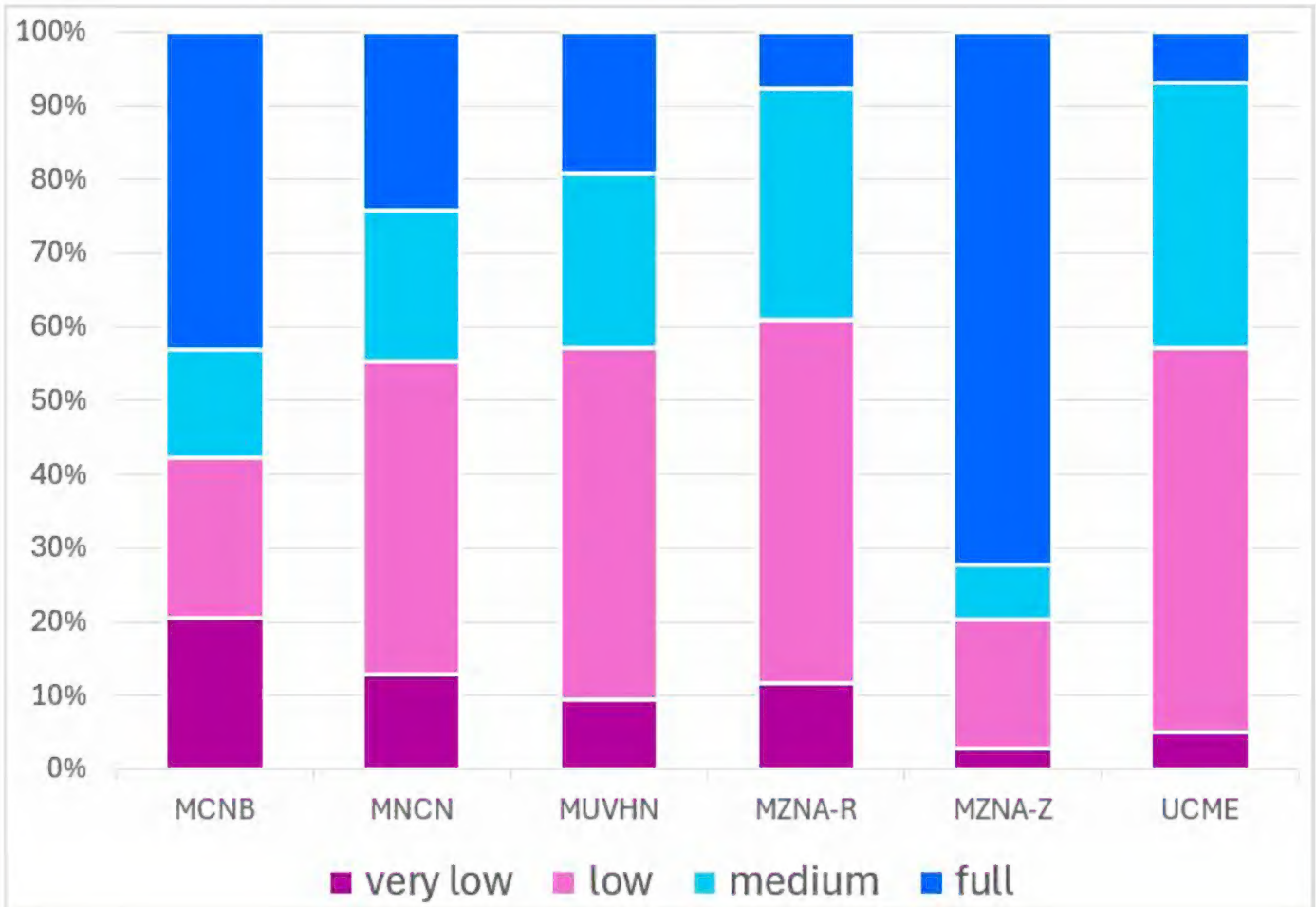

Figure 3.

Attribution rates of collectors. Pinks: poor attribution; blues: good attribution. © 2025 A.H. Ariño, CC BY-SA 4.0

Keywords

data capture, data quality, attribution, collector names, disambiguation

Presenting author

Arturo H. Ariño

Presented at

Living Data 2025

Funding program

INC-STEP project has received funding from the EU Horizon Europe programme within the framework of the TETTRIs project (GA Nr 101081903).

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Groom Q, Bräuchler C, Cubey R, et al. (2022) The disambiguation of people names in biological collections. Biodiversity Data Journal 10 https://doi.org/10.3897/bdj.10.e86089
- Hill AW, Otegui J, Ariño AH, Guralnick RP (2010) GBIF position paper on future directions and recommendations for enhancing fitness-for-use across the GBIF network, version 1.0. Global Biodiversity Information Facility, 30 pp.

Supplementary material

Suppl. material 1: Tables S1-S3 doi

Authors: Arturo H. Ariño
Data type: Procedures
Brief description: S1: Collections description. S2: Disambiguation/Attribution sequence of procedures. S3: Identity qualification levels.
Download file (94.22 kb)